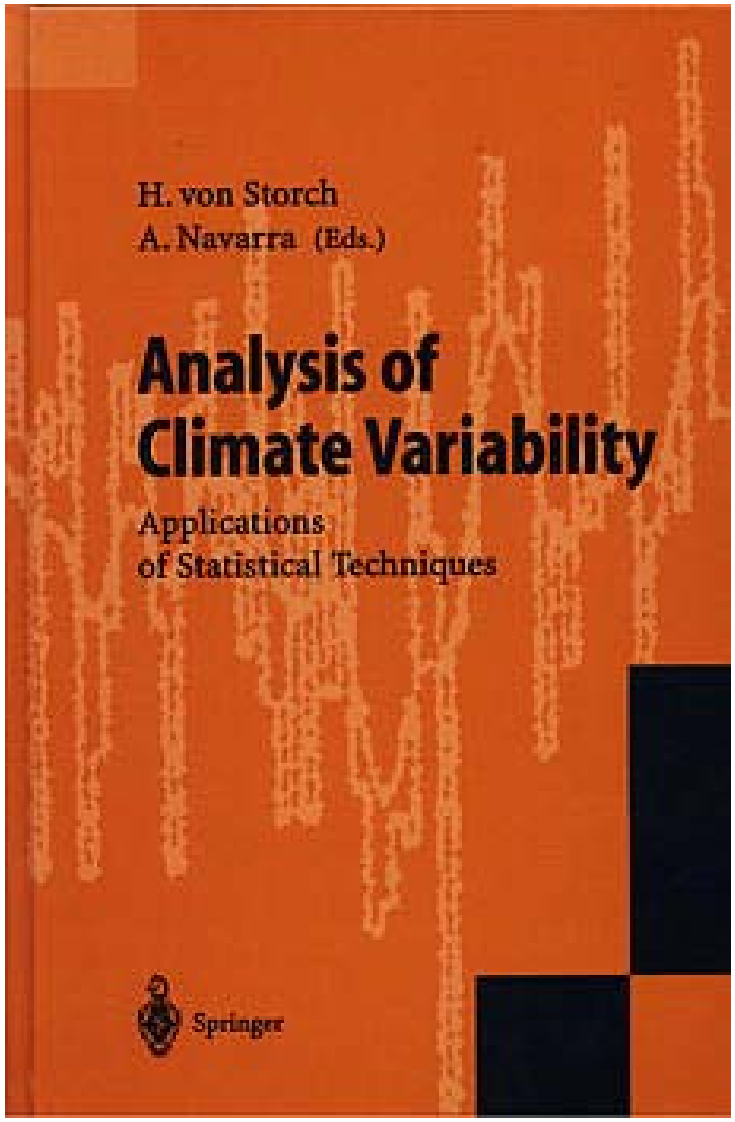
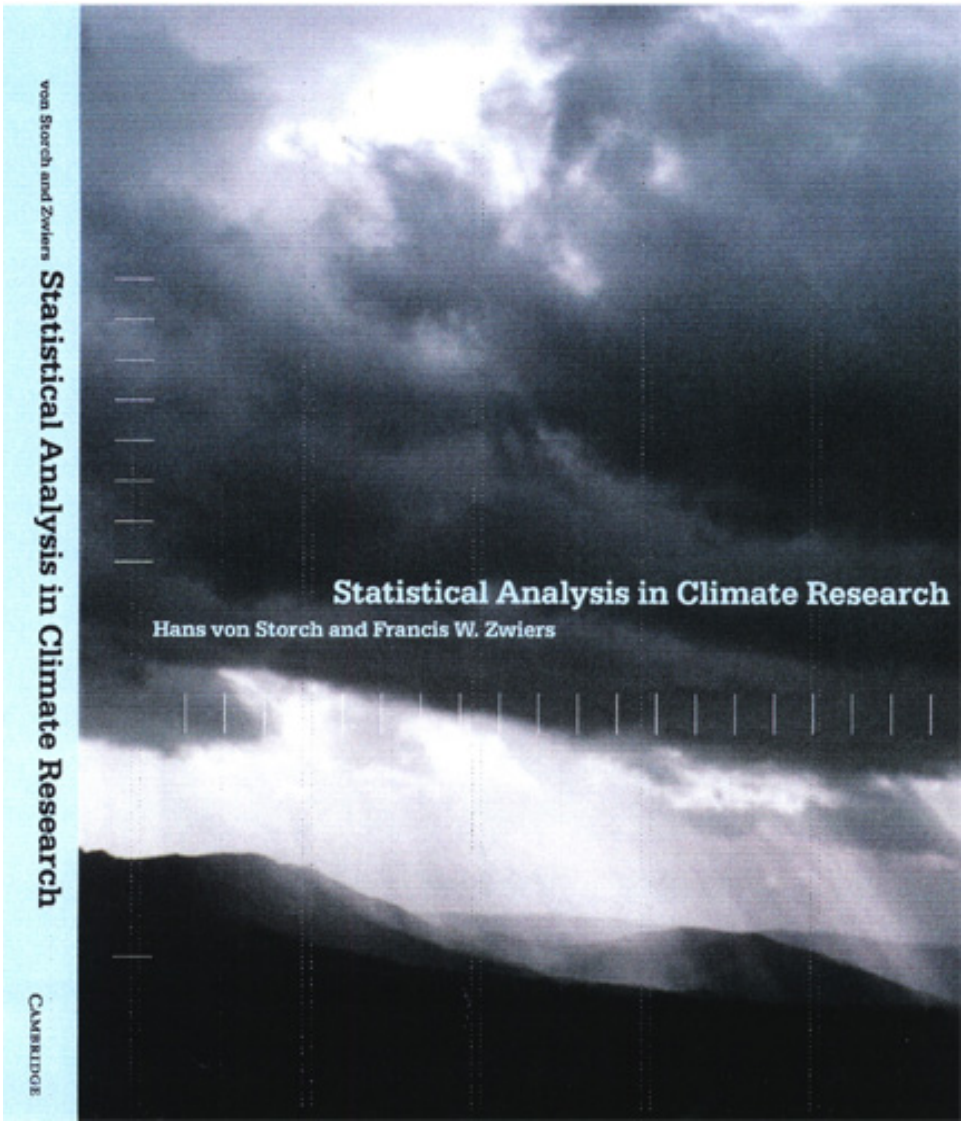
Ingredients of statistical hypothesis testing – and their significance

Hans von Storch, Institute of Coastal Research, Helmholtz-Zentrum Geesthacht, Germany

13 January 2016, New Orleans; 23rd Conference on Probability and Statistics in the Atmospheric Sciences;
Session: Best Practices for Statistical Inference in the Atmospheric Sciences


H. von Storch
A. Navarra (Eds.)
Analysis of
Climate Variability
Applications
of Statistical Techniques
Springer


von Storch and Zwiers Statistical Analysis in Climate Research
CAMBRIDGE
Statistical Analysis in Climate Research
Hans von Storch and Francis W. Zwiers

Before applying a formula for asserting the “significance” of an observation, some preparations are needed, mostly the formulation of a null-hypothesis, which (in most cases) is hoped to be rejected. This null-hypothesis includes the assumption that a variable considered would be random and governed by a certain probability distribution. When we want to find out if an “observation” of interest is in contradiction to the null-hypothesis, then we determine if this observation is in the tails of the probability distribution used in the null-hypothesis. In that case, we call the tested observation as “significant”, or more precisely, “significantly inconsistent with the null-hypothesis”. For doing so, we must in principle be able to identify all possible outcomes of the random variable. This is a non-trivial assumption; if there is a group of outcomes which are for whatever reason not accessible, we reject the null-hypothesis “a member of this group” too often. When we sample all admissible outcomes, we may estimate the probability distribution, and the quality of the estimation process is taken into account when conducting the hypothesis test. But if the sampling process is biased in some way, the uncertainty of the estimation may be underestimated. This is the case, when data are (even weakly) serially correlated, as is almost always the case in climatic applications. The technical part of the testing, namely the calculation of the measure of consistency (the “test statistic”) is in most cases simple, once the probability distribution is known or can be generated through Monte-Carlo simulation.
A very common problem is that of the “Mexican Hat”, namely that the formulation of the null-hypothesis is done after it the variable to be tested is known to be a rare outcome; also the issue of multiple tests is not always taken care of sufficiently. Another problem is that the word “significance”, which is used for indicating that the null-hypothesis is unlikely to apply to the tested observation, is understood in its colloquial meaning, namely that the inconsistency is relevant, even if there is no such link. In the presentation, the general principle of hypothesis testing is worked out, the assumptions are made explicit and examples of disregarding these assumptions discussed.
30 min

Frequentists' approach for determining consistency of data and assumptions

- Consider a variable **X**, which has some properties, and an observation **a.**
- Question: Is **a** consistent with **X**?
- For deciding, we consider **X** a random variable, i.e., an infinite number of samples x of **X** can be drawn. We know which values **X** can take on, and we know the probabilities of all possible outcomes x , i.e., P(x).
- A subset ς of x's is determined so that $\int_{\varsigma} P(x)dx = \alpha$, with a small number α.
- If a $\in\varsigma$, then the probability for any sample drawn from **X** to be equal (close) to **a** is less than α.
- If we have chosen α "sufficiently small", which is an entirely subjective judgement, we decide to consider **a** to not to be drawn from **X**, or in other words: **a** *is significantly different from* **X**.

When …? – global models … 1970s

The first literature demonstrating the need for testing the results of experiments with simulation models was ..

- Chervin, R. M., Gates, W. L. and Schneider, S. H. 1974. The effect of time averaging on the noise level of climatological statistics generated by atmospheric general circulation models. *J. Atmos. Sci.* 31, 2216–2219.
- Chervin, R. M. and Schneider, S. H. 1976a. A study of the response of NCAR GCM climatological statistics to random perturbations: estimating noise levels. *J. Atmos. Sci.* 33, 391–404.
- Chervin, R. M. and Schneider, S. H. 1976. On determining the statistical significance of climate experiments with general circulation models. *J. Atmos. Sci.* 33, 405–412
- Laurmann, J.A, and W.L. Gates, 1977: Statistical considerations in the evaluation of climatic experiments with atmospheric general circulation models, J. Atmos. Sci., 34: 1187-1199

t-test

Usually the t-test is used to determine if a number of sampled data contradict the ***hypothesis that the expectation, or population mean of the data would be zero.***

- We assume a normally distributed random variable **Y** with an expectation (mean) $\mu = 0$ and standard σ
- We repeat the random variable n-times, labelled $\mathbf{Y}_1 \ldots \mathbf{Y}_n$ – any $\mathbf{Y}_i$ generates realizations independent of all other $\mathbf{Y}_j$. All possible outcomes of **Y** may emerge as realizations, with a probability given by the distribution of **Y**

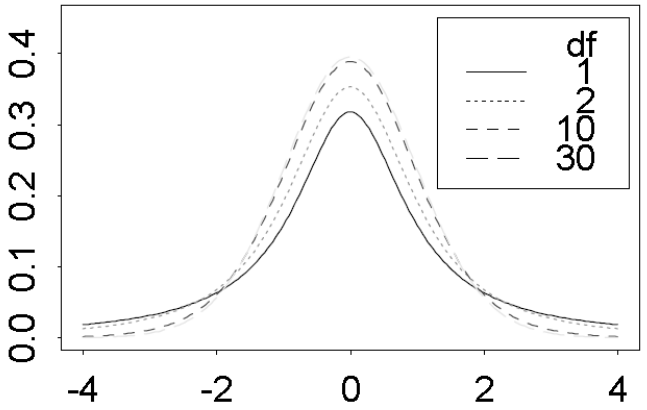


Figure 2.6: *Probability density functions for $t(k)$ random variables with 1, 2, 10, and 30 degrees of freedom.*

Then, we form the **sample mean X** $= \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i$ and the **sample variance** $\mathbf{S}^2 = \frac{\mathbf{1}}{\boldsymbol{n-1}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\mathbf{Y}_i - \mathbf{X})^2$

Then, t = $\mathbf{X}/(\mathbf{S}\sqrt{1/n})$ is a random variable, which is described by the t-distribution with n-degrees of freedom.

If we have a sample of n values $\mathbf{y}_1 \ldots \mathbf{y}_n$, which have been sampled independently and identically (from the same **Y**), then a “loo large” or “too small” t-value, this is considered evidence that the

expectation of **X** = expectation of **Y** = $\mu \neq 0$

Probability for erroneously rejecting a null hypothesis is α

- In many case, an α of 5% is chosen (social inertia).
- Thus, when I do 1000 tests, and in all cases the null hypothesis is true, I must, on average, iin 50 cases reject the null hypothesis – erroneously.
- If I do not so, my test is false.
- If all decisions are independent, the number of such false decisions is binomially distributed.
- But decisions are often not independent, in particular, when a field of locations or of variables is screened (multiplicity) $\rightarrow$ see later.

Pitfall 1 – Choose ς so that it includes **a** .

- Mexican Hat – a unique stone formation – 4 stones in vertical order reminding on a Mexican hat. This is **a**.
- Hypothesis: It is drawn from the ensemble **X** of natural formations
- We determine the frequency of formations like **a** in **X** by sampling 1 million stone formations. Since a is unique, this frequency is 1/million.
- With ς = {like **a**}, we find $P(\varsigma) \approx 10^{-6}$, and we conclude $\mathbf{a} \notin \varsigma$, or …
- **… the Mexican Hat is significantly different from natural formations.**
- By considering your finger print, I can demonstrate that you (all of you) are significantly non-human.

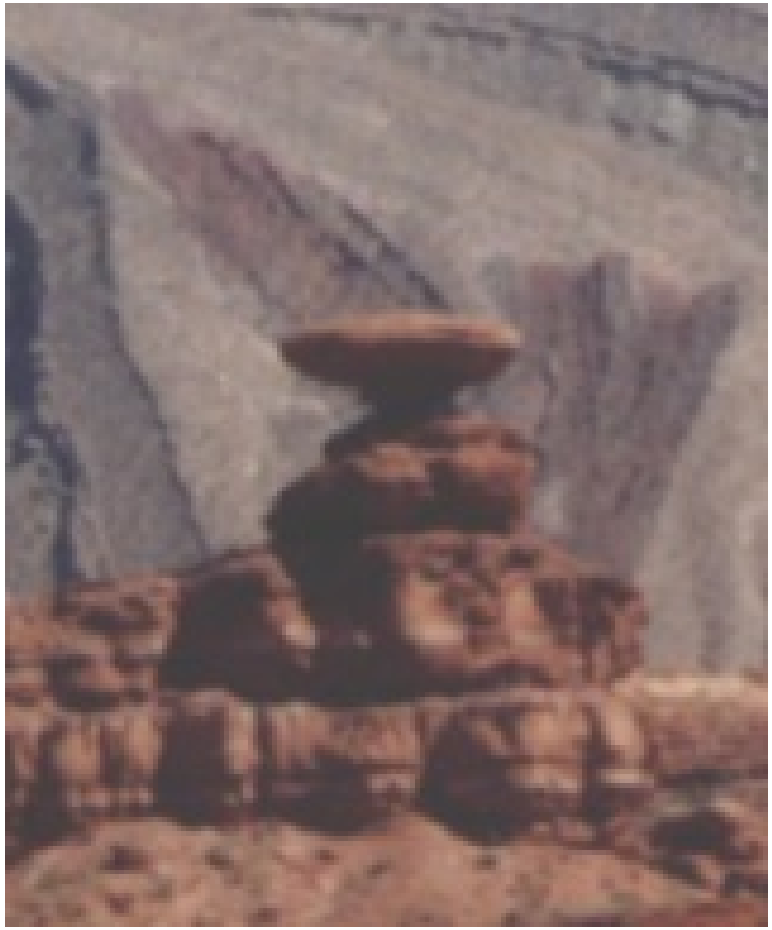


Figure 6.8: *Creation of the Mexican Hat: Null hypothesis correctly rejected!*

More general (in 1-d):

Determine a small ε so that $\int_{\mathbf{a}-\varepsilon}^{\mathbf{a}+\varepsilon} P(x)dx = 1-\alpha$ so that $\varsigma = [\mathbf{a}-\varepsilon, \mathbf{a}+\varepsilon]$ and $\mathbf{a} \in \varsigma$. All **a** are declared "significantly different from **X"**.

To make sense, the choice of the critical domain ς must be done without knowing the value of **a.**

We can define a critical domain ς by asking for a region with low probability densities (e.g., for "abnormally" large or small values), or we can ask for a region, which seems suitable to focus on because of physical insight or prior independent analysis (such as "rare and positive").

When dealing with multivariate phenomena, we have much more choices, because testing cannot be done with many degrees of freedom when noteworthy ***power*** is asked for.

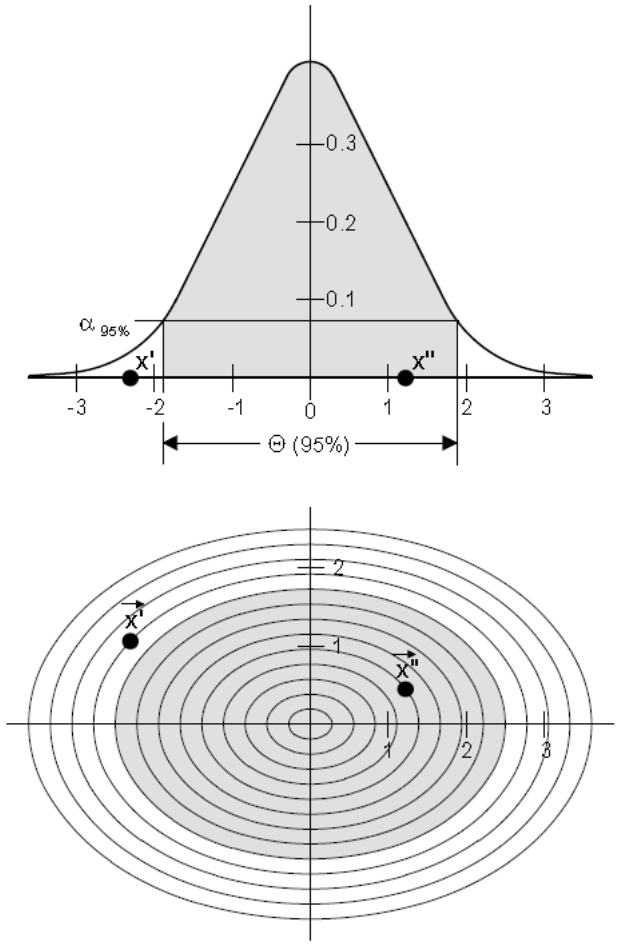


Figure 6.1: *Schematic diagrams illustrating the domains for which the null hypothesis '$\vec{\mathbf{x}}$ is drawn from $\vec{\mathbf{X}}$' is accepted. The shaded area represents the non-rejection region* $\Theta(95\%) = \{\vec{x}: f(\vec{x}) \geq \alpha_{95\%}\}$ *(a) univariate distribution; (b) bivariate distribution. The points* $\mathbf{x}'$ *and* $\vec{\mathbf{x}}'$ *are examples of realizations of the sampling process that provide evidence contrary to the null hypothesis, whereas the realizations* $\mathbf{x}''$ *and* $\vec{\mathbf{x}}''$ *are consistent with the null hypothesis [396].*

How to determine a dimensionally reduced ς in a multivariate set--up?

- Use part of the phase space where the dynamics are concentrated (e.g., given by EOFs)

- Use physical insight, what would constitute evidence against an **X**-regime

- If you can multiply generate **a** , use a first **a** for determining ς, and then draw another independent **a** to conduct the test.

However, in most cases of climate studies, we cannot draw multiple independent **a**'s from the observed record. Instead because of earlier studies, maybe by others, we already know if an event is "rare" or not. Because of (long-)memory in the climate system, the waiting time needed for the next independent realization of **a** may be very long. In that case we are in a Mexican Hat situation.

In most cases, when we deal with model simulations, we can generate multiple, independent **a**'s.

Pitfall 2: “Significance of climate change scenarios”

- There have been some attempts to qualify changes of some variables in climate change scenario simulations as “significant”.
- The problem is **X =** *outcome of climate change scenario simulation* may hardly be considered a random variable, which may be sampled such that “All possible outcomes of **X** may emerge as realizations, with a probability given by the distribution of **X**”.
- We may want to limit **X** to simulations dealing with a specific emission path, say a specific emission scenario used by CMIP.
- Can we describe “all possible outcomes”? What is the set of all (admissible) scenario simulations?
- Obviously, we cannot describe all possible outcomes, as we cannot say which models are “good enough”, and which unavailable models would be good enough but rather different from available ones.

von Storch, H. and F.W. Zwiers, 2013: Testing ensembles of climate change scenarios for "statistical significance". *Climatic Change* 117: 1-9 DOI: 10.1007/s10584-012-0551-0

Significance of scenarios …

- Thus, when we consider all “possible and admissible climate change scenario simulations” as **X**, we speak about an undefined set-up. A sampling satisfying “All possible outcomes of **X** may emerge as realizations, with a probability given by the distribution of **X**” is impossible;
- **Thus statistical testing of the hypothesis “scenario simulations using emission Scenario B1 indicate no positive shift of extreme rainfall amounts” is not possible.**
- What can be done, is limiting all simulations to a **specific model** (specific version and set-up), for which all possible pathways may be generated through variations of initial values and of some parameters. Then, significance of the scenarios can be established for that specific model – which is much less than “all scenario simulations”.
- If such an assessment is interesting, is another issue. Whenever the model is replaced by a new version, the testing needs to be repeated. Other models may show contradicting “significant” changes.

Pitfall 3: Serially dependent sampling

- In case of the test of the mean, we can derive the probability distribution of $\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i$ if the null-hypothesis is valid and if sampling $\mathbf{Y}_i$ generates realizations independent of all other $\mathbf{Y}_j$.
- In many cases of climate research, the latter assumption is not fulfilled
- Because of the inherent (long) memory in the Earth system.
- Even small serial dependencies leads to the association of too much weight of the data against the null-hypothesis of zero mean (liberal test.
- Using the concept of an "equivalent sample size" (using a t-distribution with modified number of degrees of freedom) helps little – when the "true" autocorrelation is used, the test becomes conservative, when an estimated autocorrelation is used, it becomes "liberal". Use "table-look-up test" by Zwiers and von Storch).

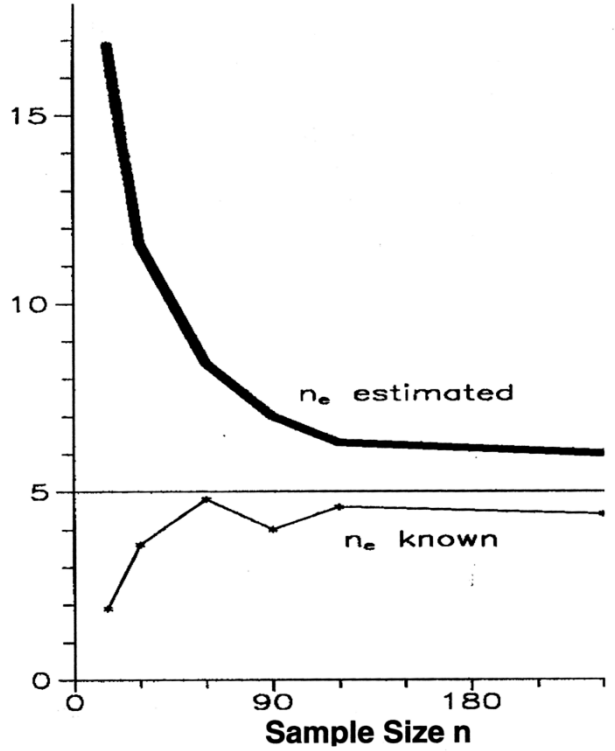


Figure 6.11: *The reject rate percentage of the one-sample t test when the observations are auto-correlated (see text). The 'equivalent sample size' n′ is given by (6.26) (thin curve) and is estimated with (6.26) (thick curve).*

Zwiers, F.W., and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. - *J. Climate* 8, 336-351

Pitfall 3: Serially dependent sampling – detecting trends

- The **Mann-Kendall test sensitive to the presence of linear trends**. Double and more false rejection rates, even for small lag-1 autocorrelations of 0.2. (Kulkarni and von Storch, 1995)
- "Prewhitening" helps in case of AR(1)-type memory in operating at the correct error-I level, but the power in correctly rejecting the null "of zero trend" is also reduced.

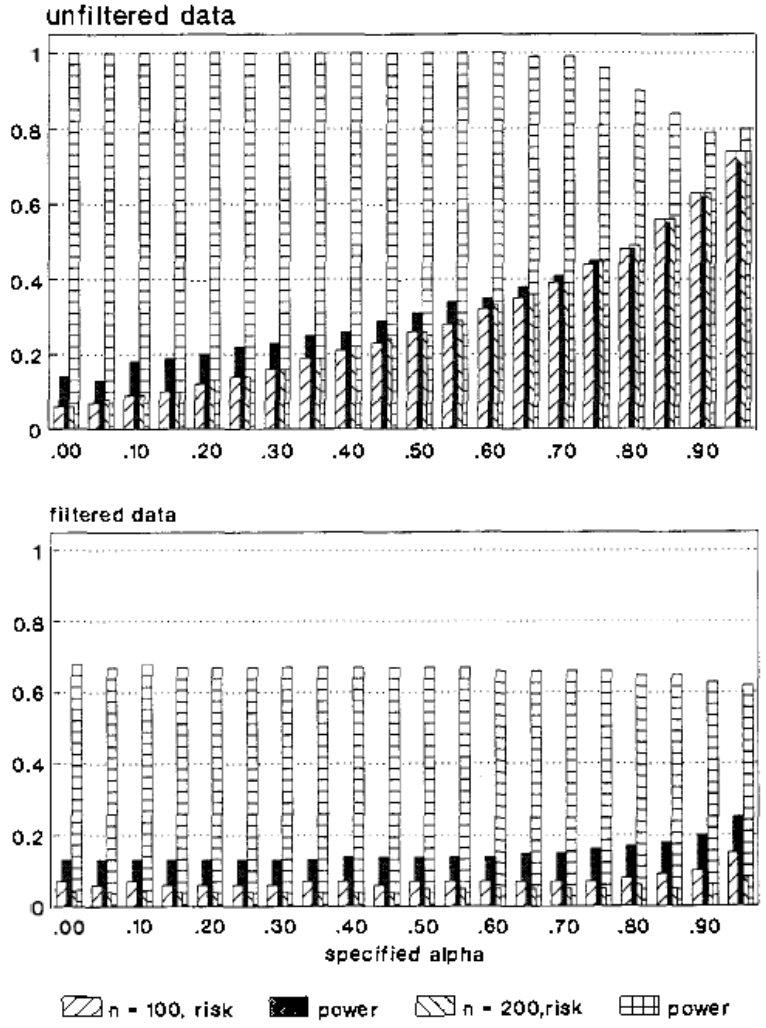


Fig. 1. Probability to reject the null hypothesis of "no trend" with the Mann-Kendall test for 1000 samples of cases without a trend ("risk") and for 1000 samples with a prescribed trend ($0.003 \times t$; "power"). Different time series lengths ($n = 100$ and $n = 200$) and different AR-coefficients α are prescribed. — Top: Results obtained with unmodified data. Bottom: Results after "prewhitening" (5) the data prior to the test.

Kulkarni, A., and H. von Storch, 1995: Monte Carlo experiments on the effect of serial correlation on the Mann-Kendall-test of trends. - *Meteor. Z* 4 NF 82-85

Pitfall 3: serially dependent sampling – detecting change points

- The Pettit-test sensitive to the presence of linear trends.
- Double and more false rejection rates, even for lag-1 autocorrelations of 0.2 (Busuioc and von Storch, 1996).
- “Prewhitening” helps in case of AR(1)-type memory in operating at the correct error-I level, but the power in correctly rejecting the null “of zero trend” is also reduced.

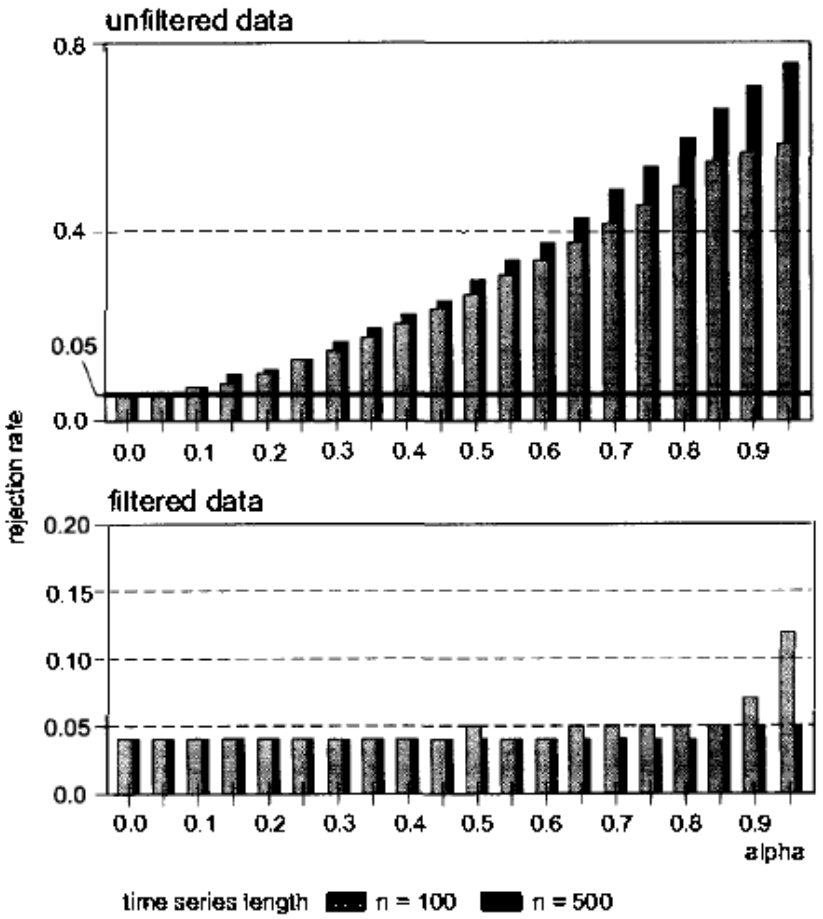


Fig. 10. Rejection rates of the Pettitt test of the null hypothesis “no change” when applied to 1000 time series of length $n = 100$ and $n = 500$ generated by an AR(1)-process (7) with prescribed α. The adopted nominal risk of the test is 5%. Top: results for unprocessed serially correlated data. Bottom: results after prewhitening the data (8).

Busuioc, A. and H. von Storch, 1996: Changes in the winter precipitation in Romania and its relation to the large-scale circulation. - *Tellus* 48A, 538-552

Pitfall 4 – Multiplicity: many “local” tests

- Often, many tests are done at the same time, e.g., when comparing an experimental simulation.
- Then multiple local test are done, and the “points” with a “local rejection” are marked.
- If the null-hypothesis of “zero mean difference” is valid at all points, at 5% of all points the null must be rejected if the test operates at the 5% level and is correctly set up.
- The number of such false rejection is itself a random variable; if the result at all points would be independent, the number of false rejections would follow a binomial distribution; however independence is in most case not given, and the distribution can be much broader (von Storch, 1982).
- Livezey and Chen (1983) have suggested a rule of thumb for deciding if “global” significance is given.

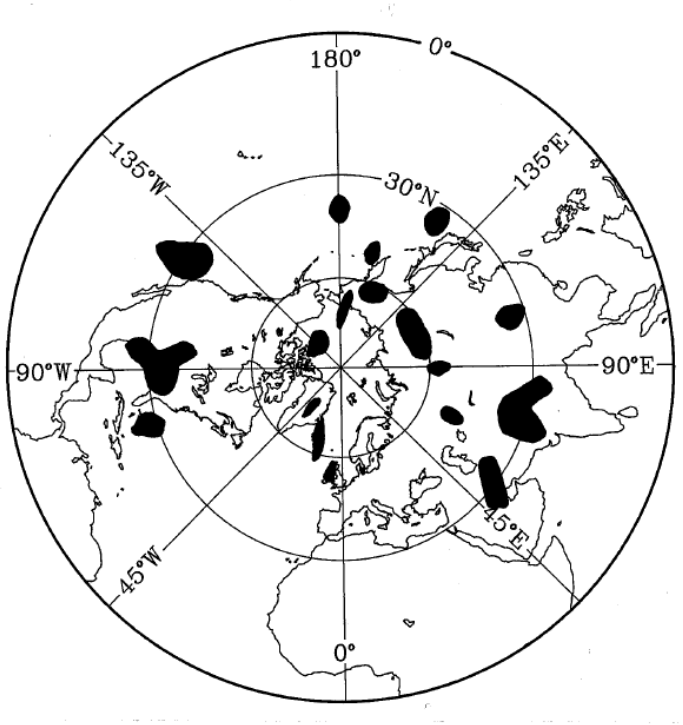


Figure 6.10: *The spatial distribution of false rejections of local null hypotheses in a Monte Carlo experiment [384].* ***All local nulls are valid.***

von Storch, H., 1982: A remark of Chervin/Schneider's algorithm to test significance of climate experiments with GCMs. *J. Atmos. Sci.* 39, 187-189

Livezey, R. E. and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.* 111, 46-59

Pitfall 5 – relevance (I) and sample size

- The probability for rejecting the null-hypothesis (power), if its actually invalid, increases with larger samples sizes.
- Thus, in case when the size of sample sizes is related to resources, then
  … a lab with limited computational resources will have fewer samples, thus less often rejection of annul-hypotheses, and will report less often "significant differences form observations" and "significant effects of an experimental change"
  … and vice versa: many samples make models more often significantly different from observed data, and seemingly more sensitive to experimental changes.
- In short:
  - poor labs have good, insensitive models,
  - rich labs have bad, sensitive models.

Pitfall 5 - Significance = relevance

- Numerical experiment – on the **effect of ocean wave dynamics on atmospheric states in the North Atlantic.**
- Regional atmospheric model, simulations with standard parameterization of ocean waves, and with explicit ocean wave dynamics.
- Measure of effect: X = daily standard deviation of SLP.
- Comparison of two 1-year simulations
- Mean ΔX shows two episodes with large spatial differences in January and July.
- Differences in January show modifications of dominant storm – physical hypothesis: storm characteristics depend on wave parameterization.

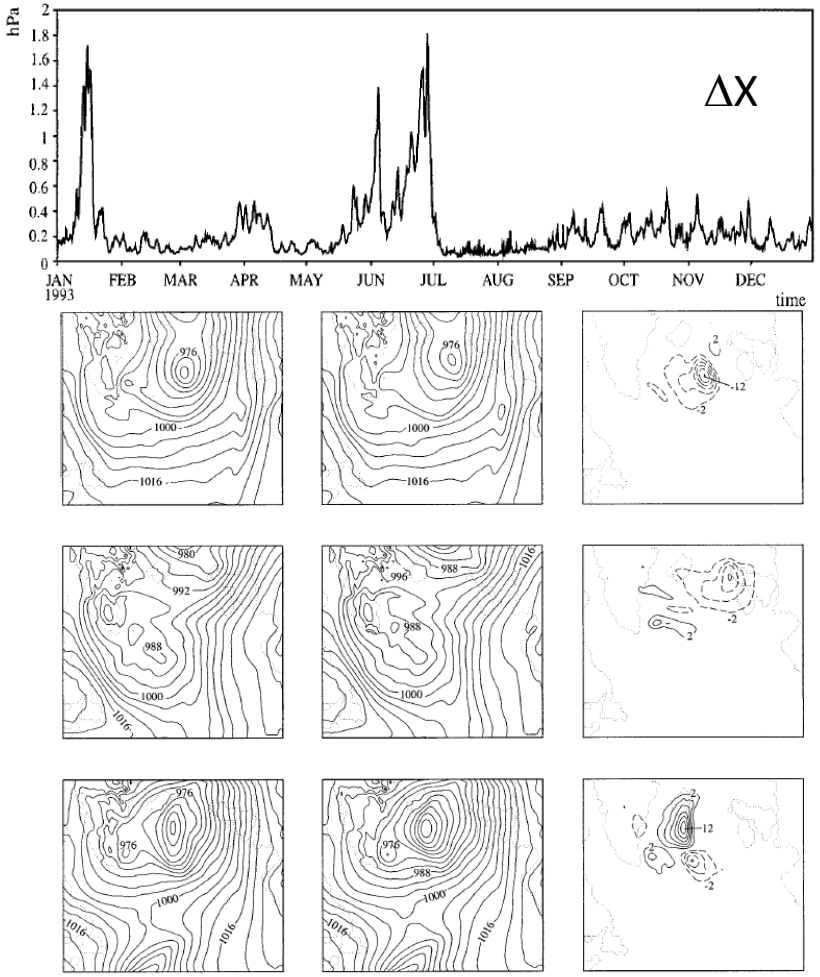


FIG. 3. SLP in hPa at 0600 UTC at 13, 14, and 15 Jan (from top to bottom) from (left) the 1-yr ESD and (middle) the 1-yr CTR simulation together with (right) the local SLP differences in hPa between both simulations.

Weisse, R., H. Heyen and H. von Storch, 2000: Sensitivity of a regional atmospheric model to a sea state dependent roughness and the need of ensemble calculations. *Mon. Wea. Rev.* 128: 3631-3642

Pitfall 5 - Significance = relevance

- Are the differences in ΔX significant? – Can we reject the null-hypothesis that the large differences are covered by the natural variability within the simulations?
- For January, 2 x 6 simulations with the same model, with standard parameterization and another with dynamic formulation of waves
- Noise levels instationary.
- When the mean differences ΔX is large, also the simulations show large ensemble-variability: synoptic situation unstable. Null not rejected.
- At the end of the month ΔX is small, and the null is rejected. Difference in employed parameterizations significant, but **signal is small and insignificant.**

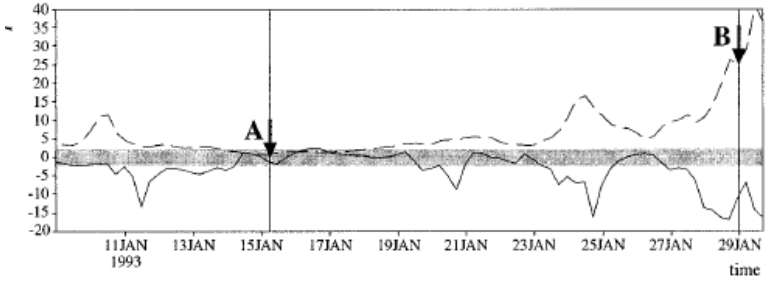
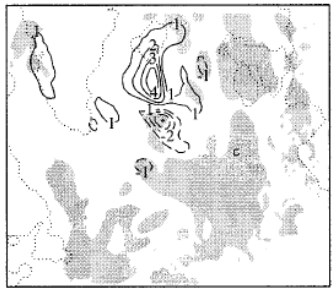
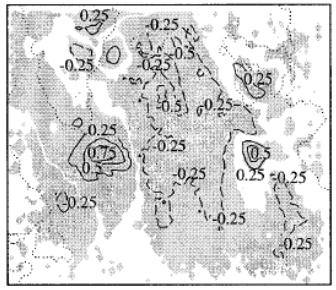


FIG. 7. The *t* statistic (5) for (solid) bias and (dashed) rmsd. The 95% confidence interval is indicated in gray. The points A and B are referred to in Fig. 8.





Mean differences ΔX at time A (top) and B (bottom) – isobares; Local significance indicated by stippling.

Weisse, R., H. Heyen and H. von Storch, 2000: Sensitivity of a regional atmospheric model to a sea state dependent roughness and the need of ensemble calculations. *Mon. Wea. Rev.* 128: 3631-3642

Take home ..

- Statistical hypothesis testing has become a standard routine in the assessment of global model simulations in climate science in the past 50 years.
- Regional modelers were late; only since about 2000 the practice is slowly entering the community.
- Here: frequentist approach – inductive conclusions constrained by some distributional and sampling assumptions.
- Example here – t-test, but a large number of approaches are in use.
- Pitfall 1 – Critical region (of rejection) chosen with knowing the signal.
- Pitfall 2 – “Significance of scenarios”
- Pitfall 3 – (Serial) dependent sampling
- Pitfall 4 – (Many) multiple tests
- Pitfall 5 – Significance = relevance